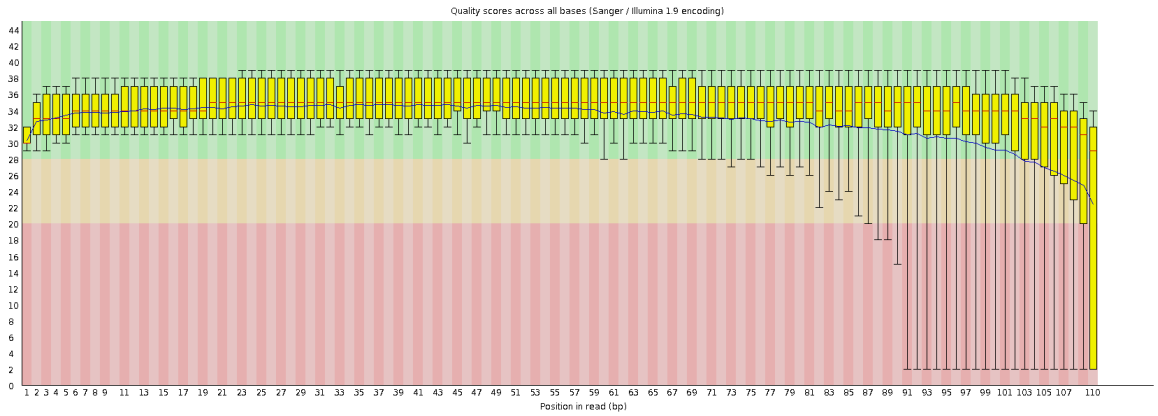
Assignment #2

Answer the following questions using the Internet, lectures from class, or the *Hosta* data set available on DropBox.

1) Outline the general steps of assembly.

2) For each step, name at least two programs that you could use to complete it.

3) Consider this FastQC output:



Would you consider this a good run? Why or why not?

4) When using a filtering script like Trimmomatic, what are we generally filtering?

5) Look at this read:

```
@Michaels-cool-machine_1:123:234:112:3215 1/N/0/88
ATTAGCCGATAGGCGAGTTTTAGAGAGAGCCGAGTATTAATTGAATAGGA
+
DDDDDEEEGFFababbbaababDFFFFFFFFFGGGGGGGGGGGGGFDDDD
```

a) What platform is it most likely from?
b) What encoding does it use for quality scores?
c) Convert the quality to a phred score.

6) What is meant by read normalization?

7) What are the benefits of read normalization? What are the drawbacks of read normalization?

8) Describe two ways to identify the coding regions of transcripts.

9) Using the three versions of the *Hosta* data set in DropBox, tell me:

a. Total number of “genes”.
b. Total number of transcripts.
c. The N50 for each.
d. What does N50 mean?

10) Blast the genes from the file data_set.txt against the three versions of the *Hosta* transcriptome. For each case, take the best hit and Blast it back against GenBank. Record in each step your best hit and any other information you think is important.

11) Annotate (using Blast) 10 sequences over 800 base pairs from the *Hosta* transcriptome.

12) How many transcripts do you have over 1000 bp? 2000 bp? Under 500 bp?